Elements of Survival Analysis

UE Statistique de la science des données

Pierre Pudlo

Aix-Marseille Université / Faculté des Sciences

Part 1. Motivation and a few definitions

What is survival analysis?

- A set of methods to analyse longitudinal data on the occurence of an event
- The event can be :
  - Medical: death, relapse, recovery
  - Social: marriage, divorce, employment or unemployment
  - Industry: time to failure of a component
  - etc.

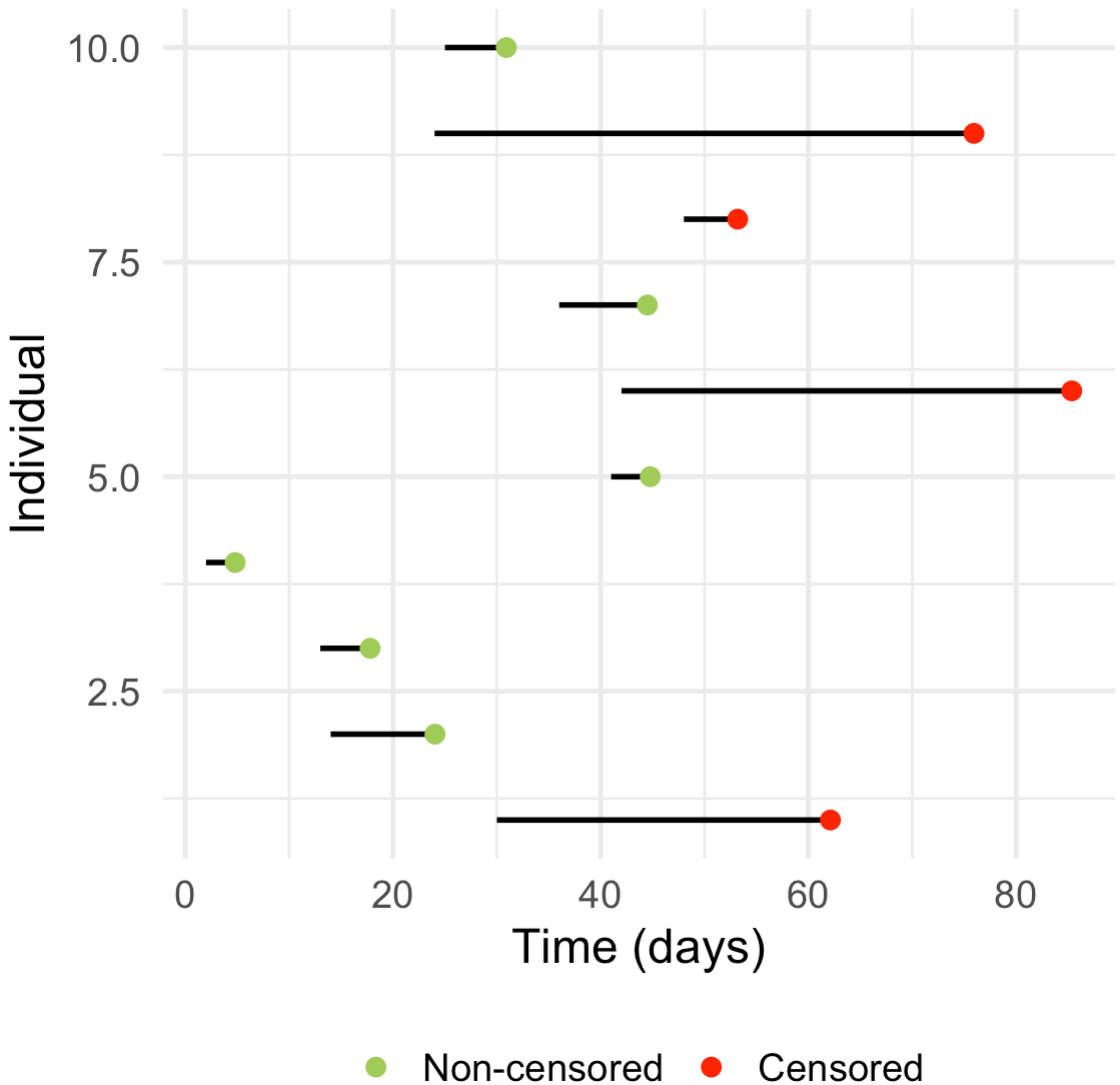
Why specific methods?

- From a machine learning perspective, time is a postive response/output
- With time comes two issues:
  - censoring: we cannot wait an infinite amount of time until the event occurs
  - vague measurement: ties (time is often day, week or month)

Vocabulary

- **Time-to-event**: the time between the entry into the study of an individual and the occurence of the event of interest
- **Right censoring**: if another event happens before the event of interest
  Examples: End of study, death by car accident, machine has been replaced, etc.

Example



- various entry dates into the study
- dataset is composed of 10 individuals
- for each individual:
  - observed duration y_i
  - non-censoring indicator δ_i

censoring is more likely if time-to-event is long

If we remove individuals that are censored from the dataset, **bias**

Data structure

For individual i, we have:

- An observed duration y_i, that is ≥ 0
- A non-censoring indicator δ_i:
  - $\delta_i = 1$ if y_i is observed
  - $\delta_i = 0$ if y_i is censored
- Maybe other covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$

True time-to-event t_i?

All we know is that:

$$\begin{cases} t_i = y_i & \text{if } \delta_i = 1 \\ t_i > y_i & \text{if } \delta_i = 0 \end{cases}$$

Random model

For individual i, we introduce random variables:

- $T_i, Y_i, \Delta_i, X_{i1}, \ldots, X_{ip}$
- C_i is the time before censoring
- $Y_i = \min(T_i, C_i) \quad ; \quad \Delta_i = \mathbf{1}\{T_i \leq C_i\}$

Assumption

Conditionnally on $\mathbf{X}_i$, we assume that T_i and C_i are independent

Density $f(t)$ of the true time-to-event

$$f(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt)}{dt}$$

- Probability **per time unit** that the event occurs at time t
- Difficult to interpret

Survival function $S(t)$ of the true time-to-event

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

where $F(t) = \mathbb{P}(T \leq t)$ is the **cumulative distribution function**

- Taking the first derivative gives the **density** as $f(t) = -S'(t)$
- Is easier to **interpret**

Hazard function $h(t)$ of the true time-to-event

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt \mid T \geq t)}{dt}$$

- Other names: instantaneous risk, intensity, force of mortality,…
- Is a probability **per time unit** that the event occurs at time t **given** that it has not occured before
- Related to the probability of the event **occuring in the next instant after time t**
- Easier to interpret than the density

Density and hazard function

When $dt \to 0$:

$$h(t)dt \sim \mathbb{P}(t \leq T < t + dt | T \geq t) = \frac{\mathbb{P}(t \leq T < t + dt)}{\mathbb{P}(T \geq t)} \sim \frac{f(t)dt}{S(t)}$$

Hence, $\quad h(t) = \dfrac{f(t)}{S(t)}.$

Since $f(t) = -S'(t)$, we have $\quad h(t) = -\dfrac{S'(t)}{S(t)} = -\dfrac{d}{dt}\Big(\log S(t)\Big)$

Interpret Hazard *vs.* Density

- At the beginning, probability that the event occurs around time t: $f(t)$
- If the event has not occured before time t, probability that it occurs just after time t: $h(t)$
- Hazard models deterioration of the system, fatigue, wear, etc.

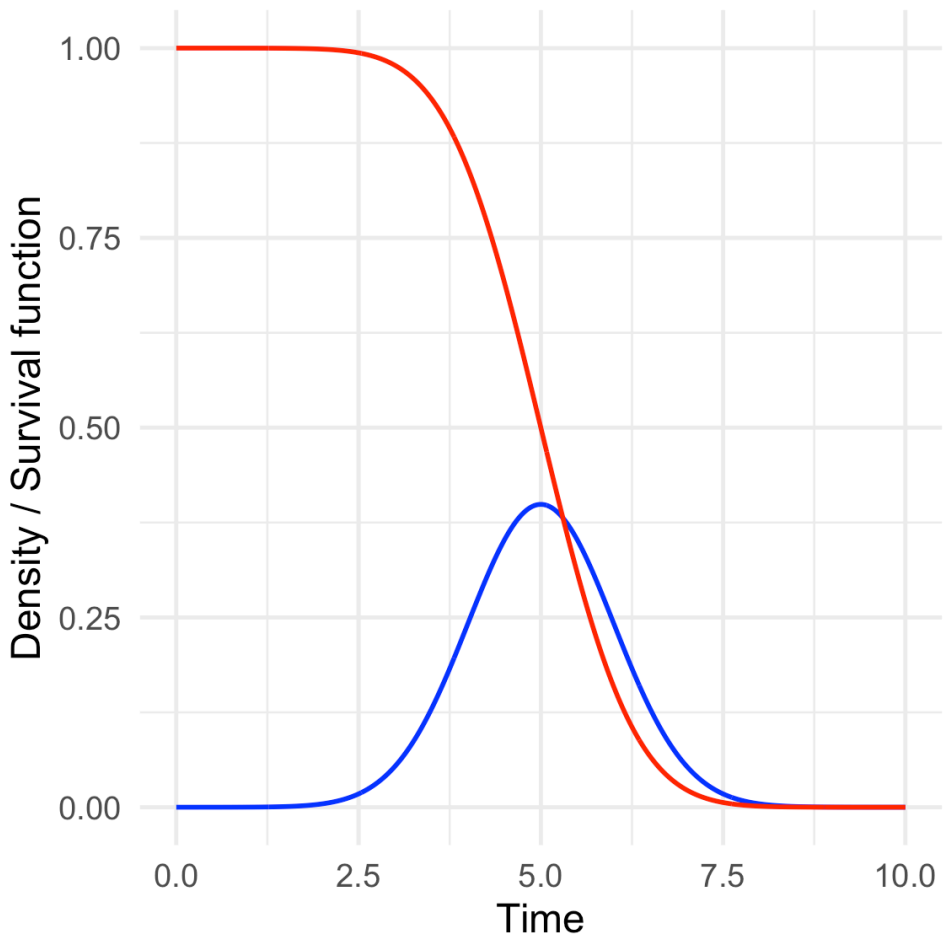
Exercise

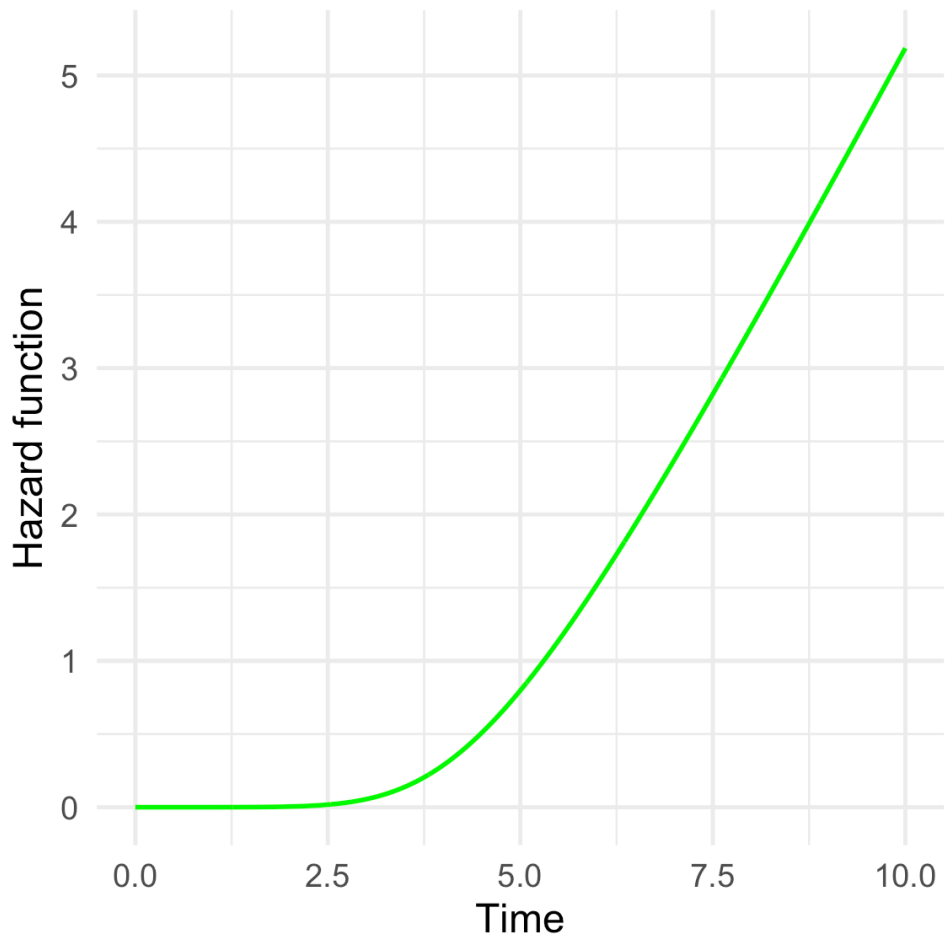
- If the hazard is constant, i.e. $h(t) = h_0$ for all $t \geq 0$, what is the density?

- When the hazard is increasing: ageing process, deterioration, wear and tear, etc.
- When the hazard is decreasing: burn-in, lapping, recovery, repair, etc.

Example 1

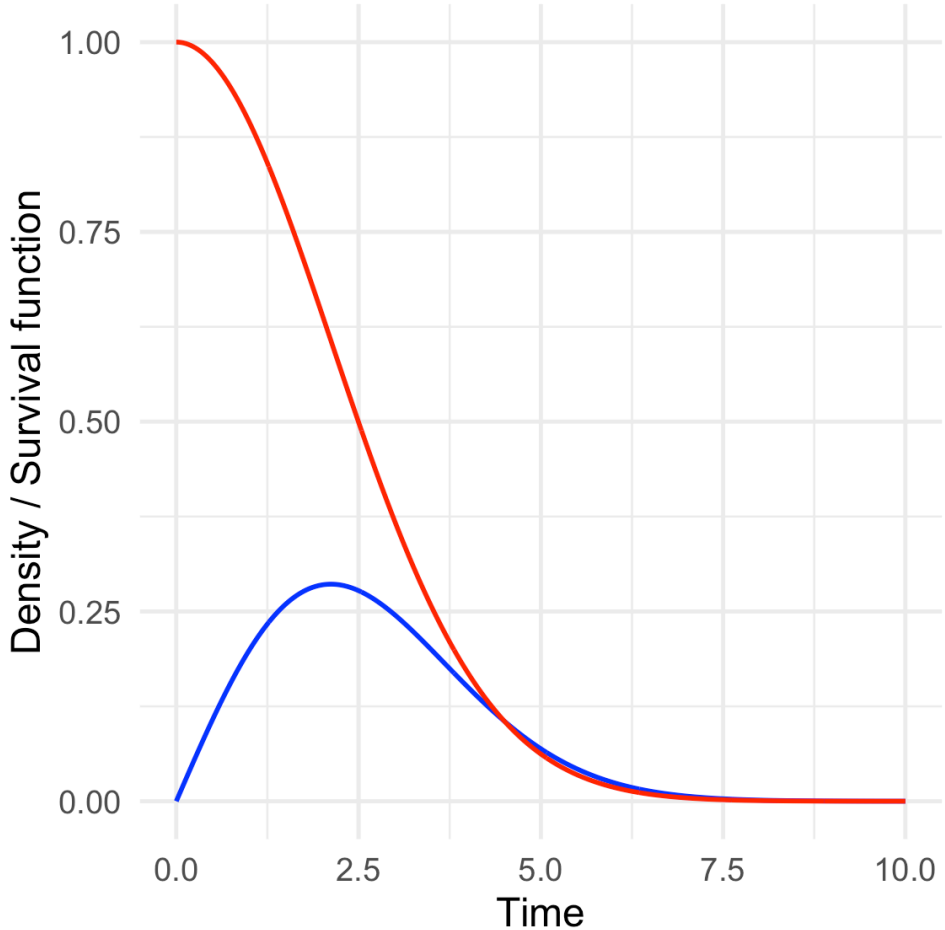
$$T \sim \mathcal{N}(5,1) \quad ; \quad f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-5)^2}{2}\right)$$

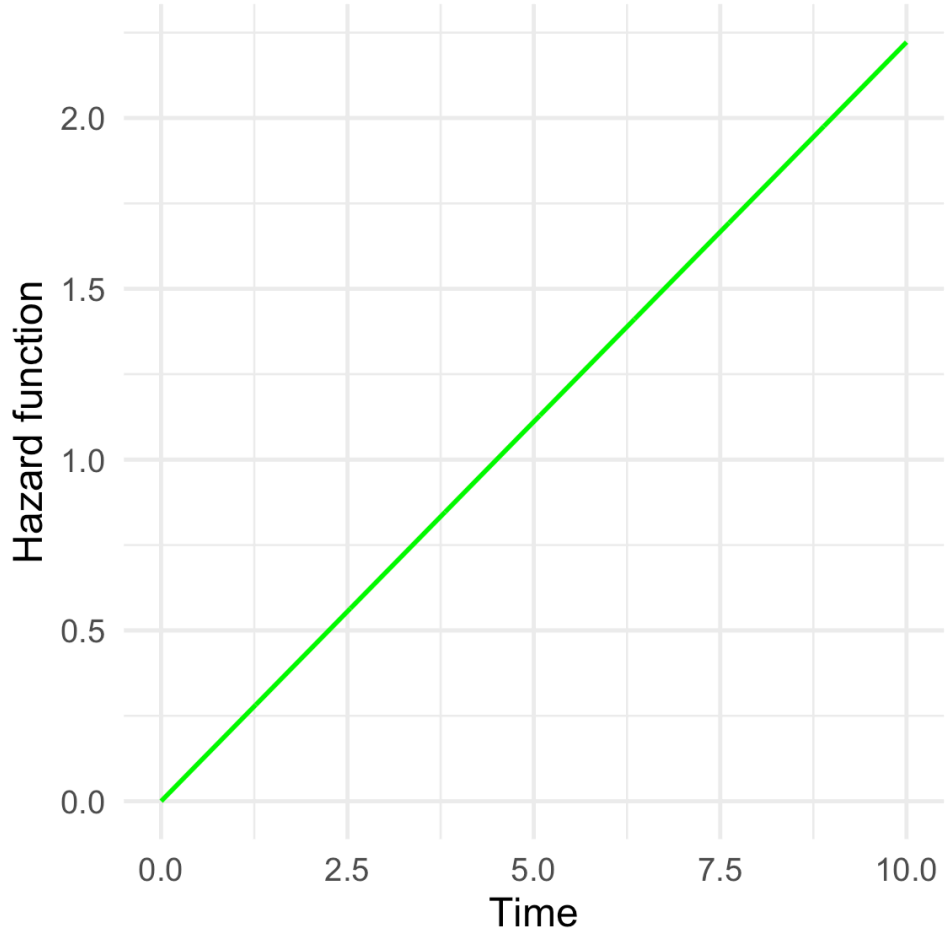




Example 2

$$T \sim \text{Weibull}(2, 3) \quad ; \quad f(t) = \frac{2}{3}\left(\frac{t}{3}\right)^{2-1} \exp\left(-\left(\frac{t}{3}\right)^{2}\right)$$


1.00
0.75
0.50
0.25
0.00
Density / Survival function
0.0
2.5
5.0
7.5
10.0
Time


2.0
1.5
1.0
0.5
0.0
Hazard function
0.0
2.5
5.0
7.5
10.0
Time

Cumulative hazard function

$$H(t) = \int_0^t h(s)ds$$

Thus,

$$H'(t) = h(t)$$

$$S(t) = \exp\left(-H(t)\right)$$

Usual distribution to model time-to-event

Exponential distribution: $T \sim \mathrm{Exp}(\lambda)$, λ : scale parameter

- $\mathbb{E}(T) = \lambda$; $S(t) = \exp(-t/\lambda)$; $h(t) = 1/\lambda$

Weibull distribution: $T \sim \mathrm{Weibull}(\gamma, \lambda)$, γ : shape, λ : scale

- $\mathbb{E}(T) = \lambda\Gamma\left(1 + \frac{1}{\gamma}\right)$, where Γ is Euler's gamma function
- $S(t) = \exp(-(t/\lambda)^{\gamma})$
- $T^{1/\gamma}$ follows an exponential distribution $\mathrm{Exp}(\lambda)$

Exercise

Compute the hazard function for the Weibull distribution

- **Log-normal distributions**: $Y = \log(T)$ follows a normal distribution
- Gamma, log-logistic, Gompertz, etc.
- **Nonparametric methods:** non hypothetis on the distribution of T

Part 2. Data without covariates

Likelihood?

- Let θ denote the unknown parameters of the model
  - $f(t|\theta)$ the density of the time-to-event
  - $S(t|\theta)$ the survival function
- Without censoring, the dataset is $t_1, \ldots, t_n$
  - the likelihood is $L(\theta) \propto \prod_{i=1}^{n} f(t_i|\theta)$
- If censoring?

Likelihood with censoring

- If the individual i is **not censored** ($\delta_i = 1$), its **contribution to the likelihood** is

$$f(y_i|\theta)$$

- If the individual i is **censored** ($\delta_i = 0$), its **contribution to the likelihood** is

$$S(y_i|\theta)$$

  since $\mathbb{P}(T_i > y_i) = S(y_i)$ is the information we have from the data

Thus, $L(\theta) \propto \prod_{i=1}^{n} \left\{ f(y_i|\theta)^{\delta_i} S(y_i|\theta)^{1-\delta_i} \right\}$

The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \left\{ \delta_i \log f(y_i|\theta) + (1-\delta_i) \log S(y_i|\theta) \right\} + \text{constant}$$

Moreover, since $S(t) = \exp(-H(t))$, we have
$f(t) = -S'(t) = H'(t)\exp(-H(t)) = h(t)\exp(-H(t))$

Hence,

$$\log S(t) = -H(t) \quad ; \quad \log f(t) = \log h(t) - H(t)$$

and

$$\ell(\theta) = \sum_{i=1}^{n} \left\{ \delta_i \log h(y_i|\theta) - H(y_i|\theta) \right\} + \text{constant}$$

Inference of parametric models

We just have to

- maximize the log-likelihood
- get confidence intervals for the parameters
- (or compute the posterior distribution if Bayesian)

Inferene of nonparametric models

Aim: infer the survival function $S(t)$ without assuming a parametric model

Classical method: Kaplan-Meier estimator

- Provides a description of the data (thus a visualisation)
- Sometimes tricky to interpret

1st example of maximum likelihood

Data: 6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316, 335*, 346* (*=censored)

$$T \sim \text{Weibull}(\gamma, \lambda) \quad \Longrightarrow \quad T^{1/\gamma} \sim \text{Exp}(\lambda)$$

	Estimates	Std.Error
γ	0.795	0.188
λ	419.164	195.414

$\mathbb{E}(T) = \lambda\Gamma\left(1 + \frac{1}{\gamma}\right)$ can be estimated by a **plug-in estimator**

```
Estimated value: 476.93
```

2nd example of maximum likelihood

Simulated data from a $\mathrm{Weibull}(4, 100)$ distribution ($n = 100$)

Data are:

```
107, 67, 95, 68, 68, 69, 95, 73, 83, 76, 103*, 93, 111, 99, 84, 85*, 120*,
91*, 80, 97, 70, 119*, 96, 35, 58, 59, 79*, 119, 81, 49* ,...
```

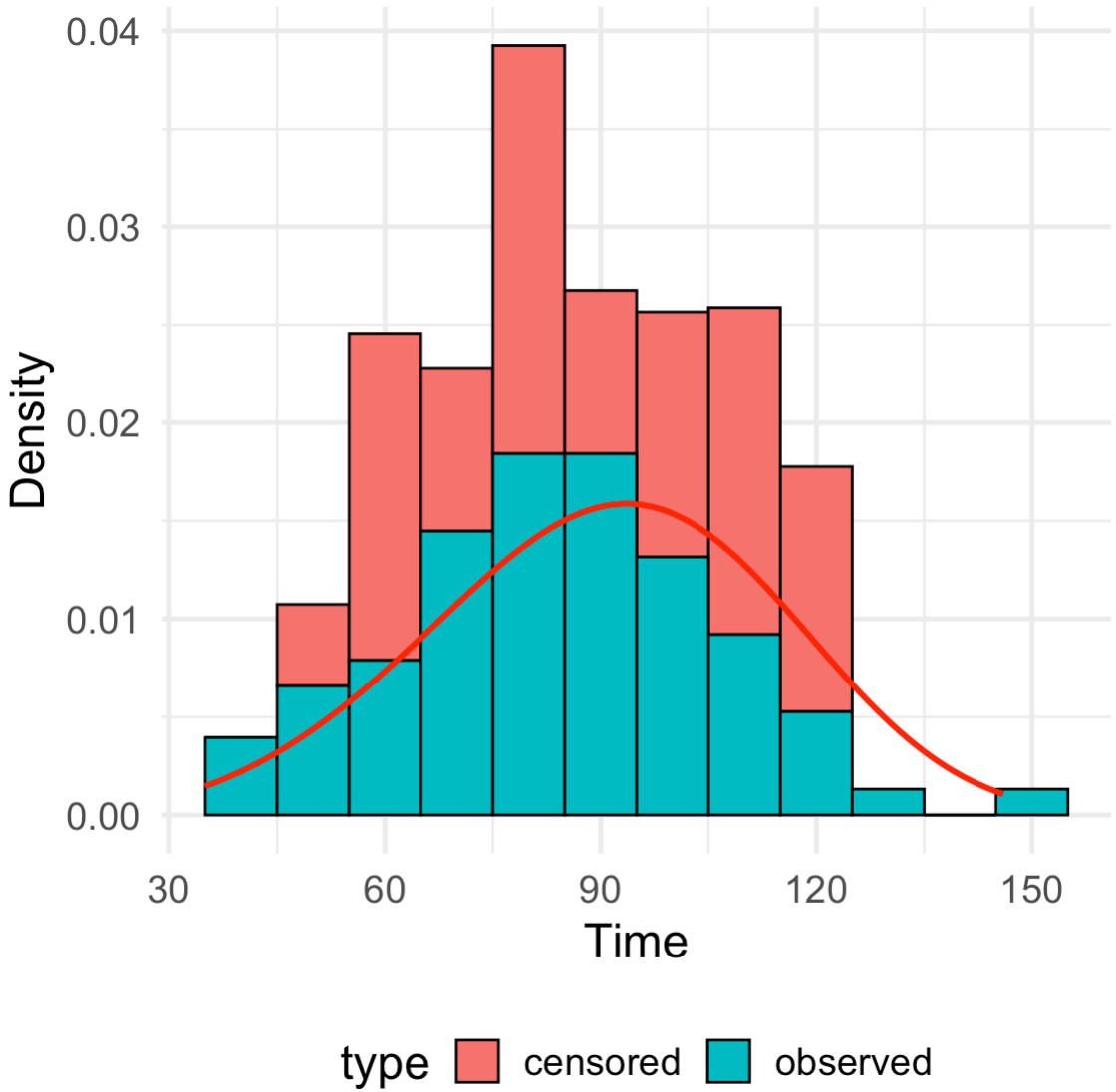
Maximum likelihood estimation of the Weibull distribution

	Estimates	Std.Error
γ	4.179	1.562
λ	99.938	2.759

Plug-in estimator of the mean, $\mathbb{E}(T)$

$$\widehat{\mathbb{E}(T)} = \widehat{\lambda}\Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right)$$

```
Estimated value: 90.81
```



- The red curve is the density of the fitted Weibull distribution


LA THEORIE
C'EST QUAND ON COMPREND TOUT
ET QUE RIEN NE MARCHE
LA PRATIQUE
C'EST QUAND TOUT MARCHE,
MAIS ON NE SAIT PAS POURQUOI
ICI
NOUS AVONS RÉUSSI LES DEUX :
RIEN NE MARCHE
ET PERSONNE NE SAIT POURQUOI

Kaplan-Meier estimator

Intialisation:

- Order the data $t_{(1)} < t_{(2)} < \cdots$ that are not censored
- Set $t_{(0)} = 0$ and start with $\widehat{S}(t_{(0)}) = 1$

Loop over $k = 1, \ldots, n$:

- We have $S(t_{(k)}) = S(t_{(k-1)})\mathbb{P}(T > t_{(k)} | T > t_{(k-1)})$
- The estimator will be a product

$$\widehat{S}(t_{(k)}) = \widehat{S}(t_{(k)})\ \mathbb{P}(T > t_{\widehat{(k)}} | T > t_{(k-1)})$$

How to estimate $\mathbb{P}(T > t_{(k)} | T > t_{(k-1)})$?

To estimate $\mathbb{P}(T > t_{(k)} | T > t_{(k-1)})$

- count $n(t_{(k)})$ the number of **individuals at risk** at time $t_{(k)}$
- count $d(t_{(k)})$ the number of **events** at time $t_{(k)}$

Then, $\widehat{\mathbb{P}(T > t_{(k)} | T > t_{(k-1)})} = \dfrac{n(t_{(k)}) - d(t_{(k)})}{n(t_{(k)})} = 1 - \dfrac{d(t_{(k)})}{n(t_{(k)})}$

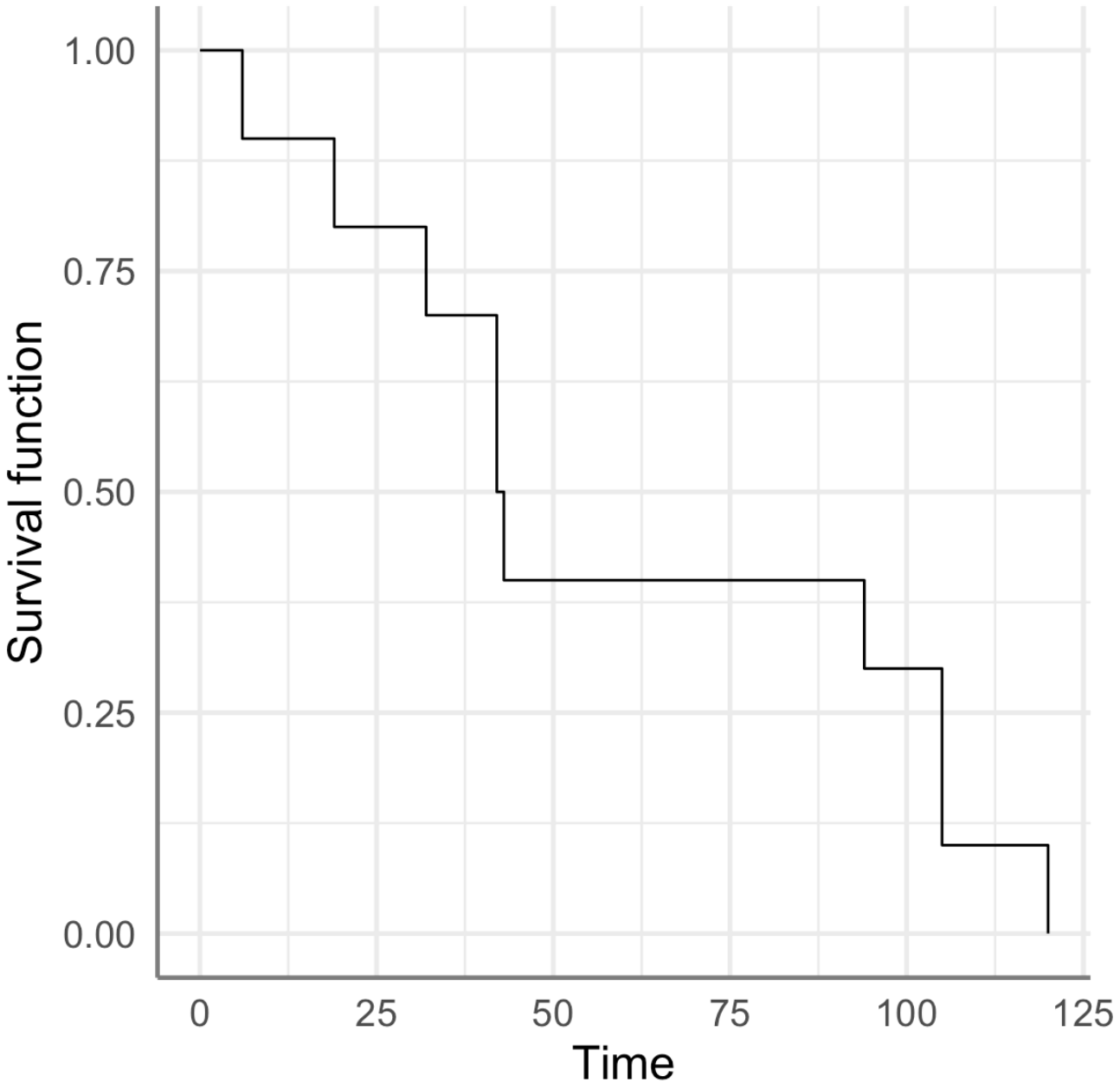
Finally, the Kaplan-Meier estimator is

$$\widehat{S}(t) = \prod_{t_{(k)} \leq t} \left(1 - \frac{d(t_{(k)})}{n(t_{(k)})} \right)$$

Example 1

data with no censoring: 6, 19, 32, 42, 42, 43, 94, 105, 105, 120

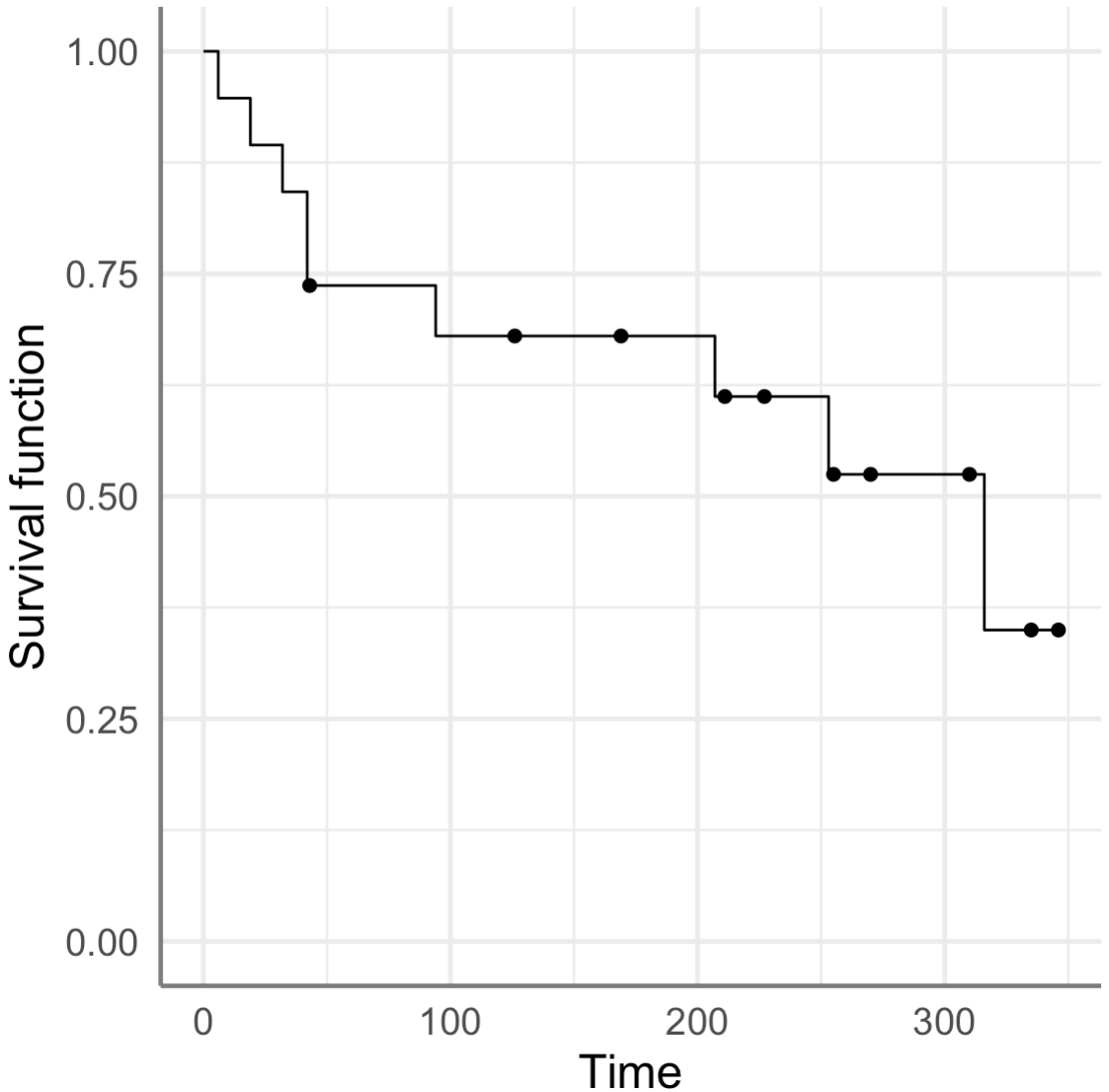
$t_{(k)}$	$n(t_{(k)})$	$d(t_{(k)})$	$1 - d(t_{(k)})/n(t_{(k)})$	$\widehat{S}(t_{(k)})$
0	-	-	-	1.0
6	10	1	0.900	0.9
19	9	1	0.889	0.8
32	8	1	0.875	0.7
42	7	2	0.714	0.5
43	5	1	0.800	0.4
94	4	1	0.750	0.3
105	3	2	0.333	0.1
120	1	1	0.000	0.0


1.00
0.75
0.50
0.25
0.00
Survival function
0
25
50
75
100
125
Time

Example 2

Data: 6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316, 335*, 346* (*=censored)

$t_{(k)}$	$n(t_{(k)})$	$d(t_{(k)})$	$1 - d(t_{(k)})/n(t_{(k)})$	$\widehat{S}(t_{(k)})$
0	-	-	-	1.000
6	19	1	0.947	0.947
19	18	1	0.944	0.895
32	17	1	0.941	0.842
42	16	2	0.875	0.737
94	13	1	0.923	0.680
207	10	1	0.900	0.612
253	7	1	0.857	0.525
316	3	1	0.667	0.350



Points are censored data

Why does the survival function decrease faster at the beginning?

Why is it not deacreasing until reaching 0?

Properties of the Kaplan-Meier estimator

- **Unbiased** estimator of the survival function $S(t)$
- **Consistent** estimator of the survival function $S(t)$

$$\widehat{S}(t) \xrightarrow[n \to \infty]{\mathbb{P}} S(t)$$

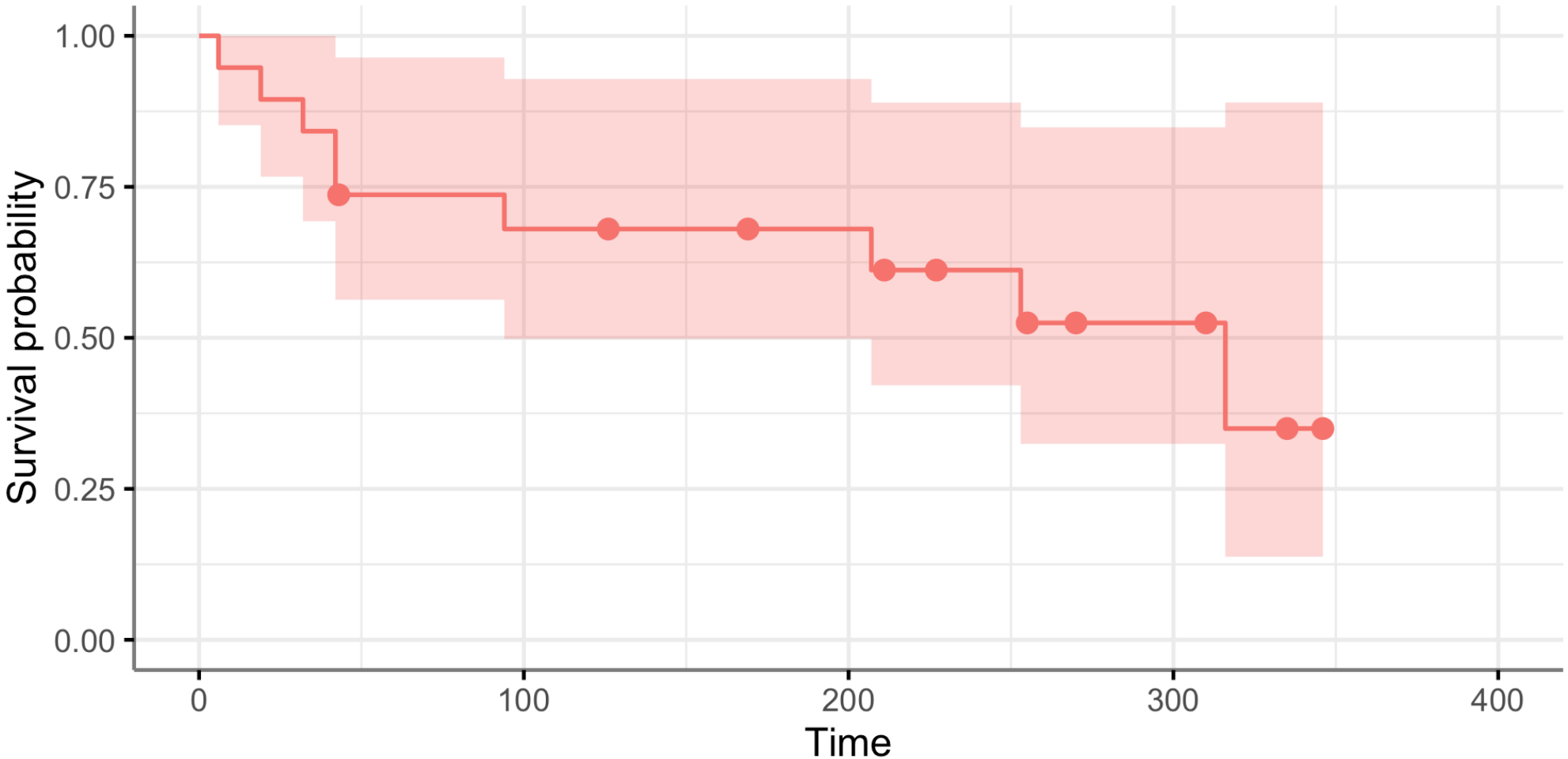
- We can compute the **variance** of the estimator and get **confidence intervals**

$$\mathbb{V}\left(\widehat{S}(t)\right) \approx \widehat{S}(t)^2 \sum_{t_{(k)} \leq t} \frac{d(t_{(k)})}{n(t_{(k)})\left(n(t_{(k)}) - d(t_{(k)})\right)}$$

$$\widehat{S}(t) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}}\left(\widehat{S}(t)\right)}$$

Plot with confidence intervals

Data: 6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316, 335*, 346* (*=censored)



Part 3. Accelerated Failure Time (AFT) models

AFT models: the idea

- 1 year of a dog life $\approx$ 7 years of a human life
- If the life of a human is the reference,
  - life of a dog is 7 times faster
  - life of a cat is 5 times faster
  - life of a turtle is 0.5 times faster
- The acceleration factor will be explained by the covariates
- Mathematically, we have, if $S_0(t)$ is the survival function of the reference individual:

$$S_{\text{dog}}(t) = S_0(7t) \quad ; \quad S_{\text{cat}}(t) = S_0(5t) \quad ; \quad S_{\text{turtle}}(t) = S_0(0.5t)$$

Exercise

If we assume that the time-to-event T_i is linked to the time-to-event of the reference $T_{0,i}$ by :

$$\log T_i = \log T_{0,i} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

then, what is the scale factor $\phi(x_{i1}, \ldots, x_{ip})$ between the reference $T_{0,i}$ and T_i, such that the survival function of T_i is

$$S_i(t) = S_0\left(t / \phi(x_{i1}, \ldots, x_{ip})\right)?$$

Answer:

$$\phi(x_{i1}, \ldots, x_{ip}) = \exp\left(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) = \prod_{j=1}^{p} \exp(\beta_j x_{ij})$$

Hence,

- covariates have a **multiplicative effect** on the time-to-event
- $\exp(\beta_j)$ is the **factor** multiplying the time-to-event for a unit increase of x_{ij}, all other things being equal:
  - $\exp(\beta_j) > 1$: the time-to-event is longer
  - $\exp(\beta_j) < 1$: the time-to-event is shorter

- The **distribution of the reference** $T_{0,i}$ is taken into a parametric model
  - log-normal: $\log T_{0,i} \sim \mathcal{N}(\beta_0, \sigma^2)$
  - Weibull: $T_{0,i} \sim \text{Weibull}(\gamma, \lambda)$
  - etc.
- If **Weibull**, then $T_{0,i} = \lambda V_{0,i}^{1/\gamma}$ where $V_{0,i} \sim \text{Exp}(1)$. Thus, $\log T_{0,i} = \log \lambda + \gamma^{-1} \log V_{0,i}$ and

$$\log T_i = \log \lambda + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma^{-1} \log V_{0,i}$$

- If **log-normal**, the only difference with a **linear model explaining** $\log T$ is censoring: $\log T_{0,i} = \beta_0 + \sigma \varepsilon_i$ and thus

$$\log T_i = \log T_{0,i} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma \varepsilon_i$$

- AFT models are parametric models, with all the pro's and con's of parametric models

Example

- T_i is the time-to-failure of a component
- X_i is the temperature of the component
- We assume that the time-to-failure is linked to the temperature by

$$\log T_i = \log T_{0,i} + \beta X_i$$

- We assume that the distribution of $T_{0,i}$ is Weibull

Estimation of AFT models

```
Call:
survreg(formula = Surv(time, status) ~ temp, data = data, dist = "weibull")
              Value Std. Error     z       p
(Intercept)  9.4709     0.5548 17.07 < 2e-16
temp        -0.0614     0.0063 -9.75 < 2e-16
Log(scale)  -1.4483     0.3038 -4.77 1.9e-06

Scale= 0.235

Weibull distribution
Loglik(model)= -36.1   Loglik(intercept only)= -55.5
    Chisq= 38.84 on 1 degrees of freedom, p= 4.6e-10
Number of Newton-Raphson Iterations: 9
n= 19
```

- `Intercept`: $\log \lambda$ of the Weibull distribution of the reference $T_{0,i}$
- `|Log(scale)|`: $1/\gamma$ of the Weibull distribution of the reference $T_{0,i}$
- `temp`: β in the model $\log T_i = \log T_{0,i} + \beta X_i$

Another example on recidvism

An experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from prison.

- `week`: week of the 1st arrest after release, or censoring time
- `arrest`: 1 if arrested, 0 if not arrested
- `fin`: 1 if the individual was in a financial assistance program, 0 otherwise
- `age`: age of the individual at the time of release
- `race`: “black” or “other”
- `wexp`: “yes” if the individual has work experience prior to incarceration, “no” otherwise
- `mar`: “married” or “not married”
- `paro`: “yes” if the individual was on parole, “no” otherwise
- `prio`: number of prior arrests
- `educ`: level of education

week	arrest	fin	age	race	wexp	mar	paro	prio	educ
20	1	no	27	black	no	not married	yes	3	3
17	1	no	18	black	no	not married	yes	8	4
25	1	no	19	other	yes	not married	yes	13	3
52	0	yes	23	black	yes	married	yes	1	5
52	0	no	19	other	yes	not married	yes	3	3
52	0	no	24	black	yes	not married	no	2	4
23	1	no	25	black	yes	married	yes	0	4
52	0	yes	21	black	yes	not married	yes	4	3

```
Call:
survreg(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio, data = Rossi)
                 Value Std. Error     z       p
(Intercept)     4.0766     0.4928  8.27 < 2e-16
finyes          0.2722     0.1380  1.97 0.04852
age             0.0407     0.0160  2.54 0.01096
raceother       0.2248     0.2202  1.02 0.30721
wexpyes         0.1066     0.1515  0.70 0.48196
marnot married -0.3113     0.2733 -1.14 0.25473
paroyes         0.0588     0.1396  0.42 0.67355
prio           -0.0658     0.0209 -3.14 0.00167
Log(scale)     -0.3391     0.0890 -3.81 0.00014

Scale= 0.712
```

Interpretation of the results

- `fin`: individuals in the financial assistance program have a **longer time-to-arrest** than those who are not

	exp(coef)	lower .95	upper .95
finyes	1.312801	1.001764	1.720412

- `age`: the time-to-arrest **increases with age**

	exp(coef)	lower .95	upper .95
age	1.041554	1.009391	1.074742

- `prio`: the time-to-arrest **decreases with the number of prior arrests**

	exp(coef)	lower .95	upper .95
prio	0.9363023	0.8986518	0.9755302

Predict?

Exercise

How to compute

- **Predicted time-to-event** for a new individual with covariates $\mathbf{x}$?
- **Predicted survival function** for a new individual with covariates $\mathbf{x}$?
- **Predicted remaining lifetime** for an observed individual that has been censored, given y_i and $\mathbf{x}_i$?

Part 4. Cox proportional hazards models

Cox model: the idea

- **Hazard function** of individual i is

$$h_i(t|\mathbf{x}_i) = h_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

- $h_0(t)$ is the **baseline hazard function** that is the same for all individuals
- $\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$ is the **hazard ratio** between two individuals with covariates $\mathbf{x}_i$ and $\mathbf{x}_{i'} = \mathbf{0}$
- Covariates have a **multiplicative effect** on the hazard function
- The **baseline hazard function** is not estimated or estimated with a nonparametric method
- This is a semi-parametric model

Estimation of the Cox model

- We fit the model by optimizing a partial likelihood. Without ties, **Cox partial likelihood** is

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{h(y_i|\mathbf{x}_i, \beta)}{\sum_{j:y_j \leq y_i} h(y_i|\mathbf{x}_j, \beta)}$$

- Cox partial likelihood does not depend on the **baseline hazard function** $h_0(t)$

  $\Longrightarrow \widehat{\beta} = \arg\max_\beta PL(\beta)$

- Cox partial likelihood depends only on the **rank** of the observed times, and not on their values

  $\Longrightarrow$ robustness of Cox models

- In presence of ties in y_i, totally ranking the data is impossible. The ranking is (partially) latent.
- Two methods to circumvent the tie problem:
  - **Breslow method**: fast but rough approximation of Cox partial likelihood
  - **Efron method**: better approximation of Cox partial likelihood but slower
- Another problem can occur when trying to optimize Cox partial likelihood: the maximum can be at $\beta_j = +\infty$ or $-\infty$ for some j
  $\Longrightarrow$ **Firth correction** to avoid this problem

Example on recidivism dataset

```
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio, data = Rossi)

  n= 432, number of events= 114

                   coef exp(coef) se(coef)      z Pr(>|z|)
finyes          -0.37942   0.68426  0.19138 -1.983  0.04742 *
age             -0.05744   0.94418  0.02200 -2.611  0.00903 **
raceother       -0.31390   0.73059  0.30799 -1.019  0.30812
wexpyes         -0.14980   0.86088  0.21222 -0.706  0.48029
marnot married  0.43370   1.54296  0.38187  1.136  0.25606
paroyes         -0.08487   0.91863  0.19576 -0.434  0.66461
prio             0.09150   1.09581  0.02865  3.194  0.00140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of the results

	exp(coef)	lower .95	upper .95
finyes	0.684	0.470	0.996
age	0.944	0.904	0.986
raceother	0.731	0.399	1.336
wexpyes	0.861	0.568	1.305
marnot married	1.543	0.730	3.261
paroyes	0.919	0.626	1.348
prio	1.096	1.036	1.159

All other factors being equal, the **risk** of recidivism is reduced by financial assistance by approximately $1 - 0.694 \approx 30\%$